> **Key Definitions**

Throughout this cheat sheet, you’ll find terms and specific statistical jargon being used. Here is a rundown of all the terms you may encounter:

1. **Variable**: In statistics, a variable is a quantity that can be measured or counted. In data analysis, a variable is typically a column in a data frame.
2. **Descriptive statistics**: Functions that summarize variables. They are also called summary statistics or aggregations.
3. **Category data**: Data that consists of discrete groups. The categories are called intervals (e.g., educational levels) if you can sort them from lowest to highest, and ordered otherwise (e.g., order of origin).
4. **Numerical data**: Data that consists of numbers (e.g., age).

---

> **Numerical Dataset—Glasses of Water**

To illustrate statistical concepts on numerical data, we’ll use a numerical variable, consisting of the volume of water in different glasses.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>How to Find It</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic mean</td>
<td>The total of the values divided by how many values there are</td>
<td>206.7 ml</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>The middle value when sorted from smallest to largest</td>
<td>180 ml</td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>The most common value</td>
<td>800 ml</td>
<td></td>
</tr>
</tbody>
</table>

---

> **Categorical Data—Trail Mix**

To illustrate statistical concepts on categorical data, we’ll be using an unordered categorical variable, consisting of different elements in a trail mix. Our categorical variable contains 4 almonds, 13 cashews, and 25 cranberries.

Counts and Proportions

Counts and proportions are measures of how much data you have. They allow you to understand how many data points belong to different categories in your data.

- A count is the number of times a data point occurs in the dataset.
- A proportion is the fraction of times a data point occurs in the dataset.

<table>
<thead>
<tr>
<th>Food category</th>
<th>Count</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almond</td>
<td>15</td>
<td>15 / 40 = 0.375</td>
</tr>
<tr>
<td>Cashew</td>
<td>13</td>
<td>13 / 40 = 0.325</td>
</tr>
<tr>
<td>Cranberry</td>
<td>25</td>
<td>25 / 40 = 0.625</td>
</tr>
</tbody>
</table>

---

**Visualizing Categorical Variables**

- **Bar plot**: One of the simplest charts to read which help in quick comparison of categorical data. One axis contains categories and the other axis represents values.
- **Stacked bar chart**: Used to compare subcategories within a categorical data. Can also be used to compare proportions.
- **Tree map chart**: 2D rectangle whose size is proportional to the value being measured and can be used to display hierarchically structured data.

---

> **Measures of Center**

Measures of center allow you to describe or summarize your data by capturing one value that describes the center of its distribution.

- **Arithmetic mean**: The total of the values divided by how many values there are.
- **Median**: The middle value when sorted from smallest to largest.
- **Mode**: The most common value.

---

> **Other Measures of Location**

There are other measures that you can use that can help better describe or summarize your data.

- **Percentile**: Cut points that divide the data into 100 intervals with the same amount of data in each interval (e.g., in the water cup example, the 10th percentile is 300 ml).
- **Quartile** refers to the group of percentiles that are not more than 25%. The first quartile is the same as the 25th percentile, which is 200 ml. The third quartile is the same as the 75th percentile, which is 500 ml.

---

**Measures of Spread**

Sometimes, rather than caring about the size of values, you care about how different they are.

- **Range**: The highest value minus the lowest value.
- **Interquartile range**: The third quartile minus the first quartile.

---

> **Correlation**

Correlation is a measure of the linear relationship between two variables. It is when one variable goes up, does the other variable go up or down? There are several algorithms to calculate correlation, but it is always a score between -1 and 1.

For two variables, X and Y, correlation has the following interpretation:

- **Correlation score** | Interpretation
  -1 | When X increases, Y decreases. Scatter plot forms a perfect straight line with negative slope.
  0 | Between -1 and 0: When X increases, Y decreases.
  1 | Between 0 and -1: There is no linear relationship between X and Y, as the scatter plot looks like a noisy mess.
  +1 | When X increases, Y increases. Scatter plot forms a perfect straight line with positive slope.

Note that correlation does not account for nonlinear effects, so if X and Y do not have a straight-line relationship, the correlation score may not be meaningful.